Применение методов оптимизации и машинного обучения для выстраивания образовательного процесса

Рябинин Евгений
Optimization Class Project. MIPT

Введение

Оценивание вероятности поступления определенного абитуриента в высшие учебные заведения по различным показателям вступительных данных является актуальной для большинства ВУЗов ввиду особенностей правил поступления, меняющихся из года в год, а также популярности тех или иных направлений обучения. Чтобы изучить эту проблему качественнее, я решил воспользоваться открытыми данными о поступлении 400 студентов в один из международных институтов.

Данные будут использоваться для построения регрессионных моделей, в частности линейной (МНК), обобщенной линейной (распределение Бернулли). После же будут построены графики зависимости шанса поступления от определенного критерия (всего их 7). Ошибки будут сравниваться с текущим показателем шанса поступления конкретного абитуриента из таблицы посредсвом MSE и MAE. Также будет показано иерархическая древовидная структура вероятности поступления, исходя из исходных критериев поступления. Подробнее с данными можно ознакомиться здесь.

Восстановление регрессии

Задачу обучения по прецедентам при $Y = \mathbb{R}$ принято называть задачей восстановления регрессии. Основные обозначения остаются прежними. Задано пространство объектов X и множество возможных ответов Y. Существует неизвестная целевая зависимость $y^* : X \to Y$, значения которой известны только на объектах обучающей выборки $X^\ell = (x_i, y_i)_{i=1}^\ell, y_i = y^*(x_i)$. Требуется построить алгоритм $a : X \to Y$, аппроксимирующий целевую зависимость y^*.

Метод наименьших квадратов (линейная регрессия)

Пусть модель алгоритмов задана в виде параметрического семейства функций $f(x, \alpha)$, где $\alpha \in \mathbb{R}^p$ - вектор параметров модели.

Определим функционал качества аппроксимации целевой зависимости на выборке X^ℓ как сумму квадратов ошибок:

$$Q\left(\alpha, X^\ell\right) = \sum_{i=1}^{\ell} w_i \left(f\left(x_i, \alpha\right) - y_i\right)^2$$

где w_i - вес, выражающий степень важности i-го объекта.

Обучение по методу наименъиих квадратов (МНК) состоит в том, чтобы найти вектор параметров α^*, при котором достигается минимум среднего квадрата ошибки на заданной обучающей выборке X^ℓ :

$$\alpha^* = \arg\min_{\alpha \in \mathbb{R}^p} Q\left(\alpha, X^\ell\right)$$

Стандартный способ решения этой оптимизационной задачи - воспользоваться необходимым условием минимума. Если функция $f(x, \alpha)$ достаточное число раз дифференцируема по α, то в точке минимума выполняется система p уравнений относительно p неизвестных:

$$\frac{\partial Q}{\partial \alpha}\left(\alpha, X^\ell\right) = 2\sum_{i=1}^{\ell} w_i \left(f\left(x_i, \alpha\right) - y_i\right) \frac{\partial f}{\partial \alpha}\left(x_i, \alpha\right) = 0.$$

Численное решение этой системы принимается за искомый вектор параметров α^*. Метод наименьших квадратов широко используется благодаря интуитивной ясности и удобству реализации.

Обобщённая линейная модель

Обобщённая линейная модель для математ ического ожидания:

$$y_i \sim \operatorname{Exp}\left(\theta_i, \phi_i\right), \quad \theta_i = x_i^\top \alpha = g\left(\mathrm{E}y_i\right)$$

Exp - экспоненциальное семейство распределений с параметрами θ_i, ϕ_i и параметрами-фун кциями $c(\theta), h(y, \phi)$:

$$p\left(y_i \mid \theta_i, \phi_i\right) = \exp\left(\frac{y_i\theta_i - c\left(\theta_i\right)}{\phi_i} + h\left(y_i, \phi_i\right)\right)$$

Математическое ожидание и дисперсия с.в. $y_i \sim \operatorname{Exp}\left(\theta_i, \phi_i\right)$:

$$\begin{aligned} \mu_i = \mathrm{E}y_i = c'\left(\theta_i\right) \Rightarrow \theta_i = g\left(\mu_i\right) = \left[c'\right]^{-1}\left(\mu_i\right) \\ \mathrm{D}y_i = \phi_i c''\left(\theta_i\right) \end{aligned}$$

$g(\mu)$ - монотонная функция связи (link function)

Логистическая регрессия (обощенная линейная регрессия)

Распределение Бернулли, $y_i \in \{0, 1\}$: $\quad p\left(y_i \mid \mu_i\right) = \mu_i^{y_i}\left(1 - \mu_i\right)^{1 - y_i}$

$$\theta_i = g\left(\mu_i\right) = \ln\frac{\mu_i}{1 - \mu_i} \quad \mu_i = g^{-1}\left(\theta_i\right) = \frac{1}{1 + \exp\left(-\theta_i\right)} \equiv \sigma\left(\theta_i\right)$$

Свойства:

1. $y_i - 6$ бернуллиевские случайные величины с E $y_i = \mu_i$;
2. модель линейна и μ_i монотонно зависит от $\theta_i = x_i^\top \alpha$; вытекают все основные свойства логистической регрессии:
3. логарифмическая функция потерь $\ln\left(1 + \exp\left(-\tilde{y}_i x_j^\top \alpha\right)\right)$;
4. сигмоидная функция связи $p\left(y_i \mid x_i\right) = \sigma\left(\tilde{y}_i x_i^\top \alpha\right)$;
5. связь линейной модели с отношением шансов (odds ratio):

$$x_i^\top \alpha = \theta_i = \ln\frac{\mu_i}{1 - \mu_i} = \ln\frac{\mathrm{P}\left(y_i = 1 \mid x_i\right)}{\mathrm{P}\left(y_i = 0 \mid x_i\right)}.$$

Средняя квадратичная ошибка

MSE (средняя квадратичная ошибка) является самой простой и наиболее распространенной функцией потерь (loss function). Чтобы рассчитать MSE, вы берете разницу между предсказаниями вашей модели и основополагающей правдой, возводите ее в квадрат и усредняете ее по всему набору данных.

$$MSE\left(y^{\text{truse}}, y^{\text{pred}}\right) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - f\left(x_i\right)\right)^2$$

MSE отлично подходит для гарантии того, что наша обученная модель не имеет прогнозируемых выбросов с огромными ошибками, поскольку MSE придает большее значение этим ошибкам из-за квадратной части функции. Однако если наша модель делает одно очень плохое предсказание, то квадратичная часть функции увеличивает ошибку.

Средняя абсолютная ошибка

MAE (средняя абсолютная ошибка) лишь немного отличается по определению от MSE, но, что интересно, обеспечивает почти совершенно противоположные свойства. Чтобы рассчитать MAE, вы берете разницу между предсказаниями вашей модели и основополагающей правдой, применяете абсолютное значение к этой разнице и затем усредняете его по всему набору данных.

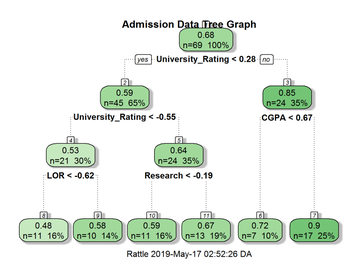
$$MAE\left(y^{\text{true}}, y^{\text{pred}}\right) = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - f\left(x_i\right)\right|$$

В отличие от MSE, мы не будем придавать слишком большой вес нашим выбросам, а наша функция потерь обеспечивает общую и даже меру того, насколько хорошо работает наша модель. Ное сли мы действительно заботимся о прогнозируемых отклонениях нашей модели, то MAE не будет столь же эффективным. Большие ошибки, возникающие из-за выбросов, в конечном итоге взвешиваются точно так же, как и более низкие ошибки.

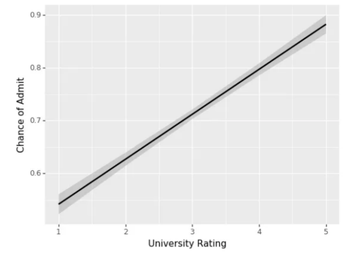
Результаты

В результате должны быть получено несколько графиков зависимости шанса поступления от определенного признака с оценкой отклонения от истинного значения (на рисунке (b) изображена подобная зависимость от University Rating).

Кроме того, после реализации нашей модели указанными методами регрессии может быть построенна иерархическая графовая структура в виде дерева принятия решений. Это поможет определить нам взаимосвязь признаков и их влияние на данные. Пример такого дерева на рисунке (a).



(a) Flower one.



(b) Flower two.

Заключение

В ходе результатов работы можно будет посмотреть на важность тех или иных признаков для нашей функции, отвечающей за "Chance of Admit а именно от чего в наибольшей степени зависит вероятность поступления и как сильно взаимосвязаны различные объекты. На рисунке ниже представлена дендограмма такой взаимосвязи, и чем раньше подключаются объекты (если смотреть с правой стороны), тем сильнее они взаимосвязаны.

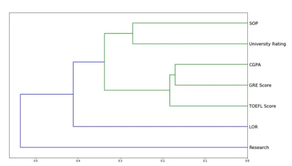


Рис. 1: Дендограмма признаков от вероятности поступления